Exploring Variables

Shannon Pileggi

STAT 217

Working with Variables

Group Exercise

An experiment regarding the physiological cost of reproduction on male fruit flies contains the following variables. Male fruit flies were randomly assigned to cohabitate with one of 5 experimental groups of female fruit flies.

`type`	Type of experimental assignment 1 = no females 2 = 1 newly pregnant female 3 = 8 newly pregnant females 4 = 1 virgin female 5 = 8 virgin females
`lifespan`	lifespan (days)
`thorax`	length of thorax (mm)

How many quantitative variables does this data set contain?

0

1

2

3

4

5

6

7

10 observations from survey results

FirstStats	gpa	target_grade	length_rel	in_rel	CP1stChoice	num_coll	num_text
No	2.500	B	48.00	No	Yes	3	10
Yes	3.000	B	36.00	No	Yes	5	100
Yes	3.389	A	24.00	Yes	No	18	100
No	3.298	B	4.00	No	No	11	30
No	3.200	A	0.25	No	No	8	100
No	2.920	B	14.00	No	No	7	600
No	3.500	A	12.00	Yes	No	6	30
Yes	2.800	A	10.00	No	Yes	13	100
No	3.470	A	23.00	No	No	13	50
No	3.050	B	6.00	No	No	11	35

`FirstStats`	first stats class?
`gpa`	GPA
`target_grade_rel`	target grade in stat 217
`length_rel`	length (in months) of longest serious relationship
`in_rel`	whether or not currently in a serious relationship
`CP1stChoice`	whether or not Cal Poly was your first choice
`num_coll`	number of colleges applied to
`num_text`	number of texts sent in a day

Summary of data produced by R

```
> summary(survey)
 FirstStats      gpa          CP1stChoice    num_coll
 No :34     Min.   :1.700   No :23      Min.   : 0.000
 Yes:33     1st Qu.:3.000   Yes:44      1st Qu.: 5.000
            Median :3.132               Median : 7.000
            Mean   :3.178               Mean   : 7.239
            3rd Qu.:3.493               3rd Qu.: 9.000
            Max.   :4.000               Max.   :18.000
            NA's   :1
```

1. How are the quantitative and categorical variables summarized differently?

2. What else do you notice?

Categorical variable

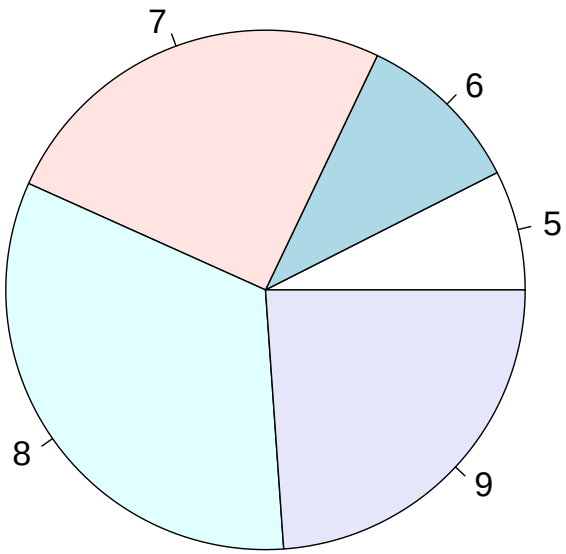
```
> addmargins(table(survey$CP1stChoice))
 No Yes Sum
 23  44  67
```

1. Identify a *statistic* that summarizes this variable.

2. Produce a visualization of this variable.

Pie charts

Opinion on the value of statistics in society
1 (completely useless) to 9 (incredible important)



Approximately what percent of students rated statistics as a 5?

1. 4%
2. 7%
3. 10%
4. 13%
5. 16%

Quantitative variable - center and variability

```
> library(mosaic)
> favstats(survey$num_coll)
 min Q1 median Q3 max     mean       sd  n missing
   0  5      7  9  18 7.238806 3.737969 67       0
```

1. Identify two measures of center, and interpet.
   -
   -
2. Identify two measures of variability, and interpret.
   -
   -

Quantitative variable - position

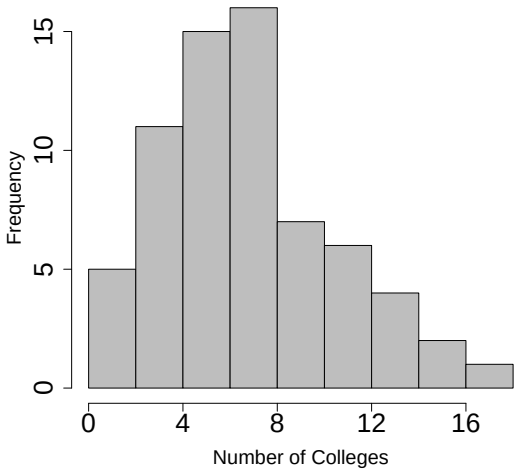
```
> library(mosaic)
> favstats(survey$num_coll)
 min Q1 median Q3 max     mean       sd  n missing
   0  5      7  9  18 7.238806 3.737969 67       0
```

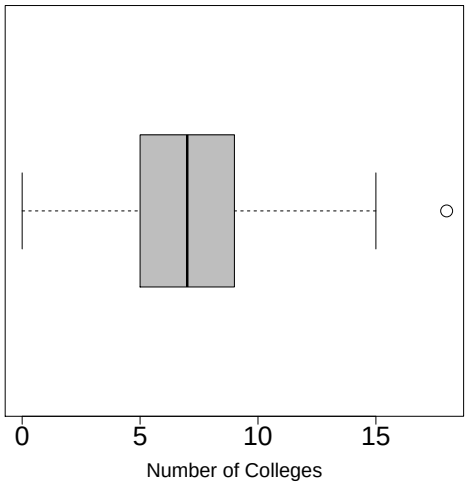
What is the value and interpretation of

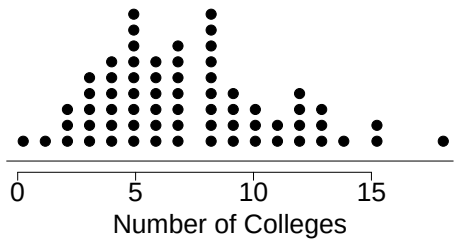
1. Q1

2. Q3

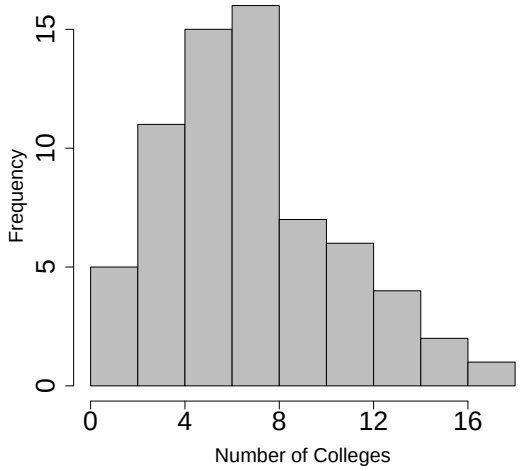
Quantitative variable - figures

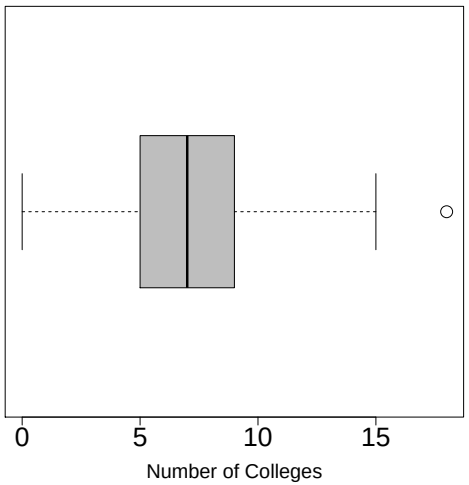


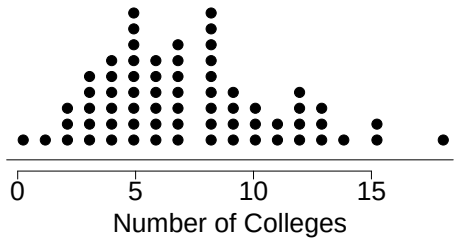




Group Exercise







True or False

1. 5 students applied to 0 colleges
2. 50% of students applied to 8 colleges or less

Group Exercise

How many variables does a histogram show the distribution of?

0

1

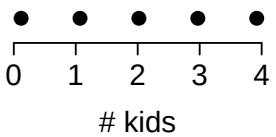
2

3

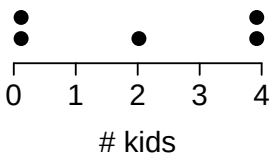
5. it depends

Suppose I asked three groups of 5 college students how many children they want to have.

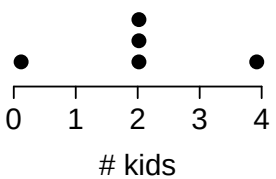
Group 1: 0, 1, 2, 3, 4



Group 2: 0, 0, 2, 4, 4



Group 3: 0, 2, 2, 2, 4

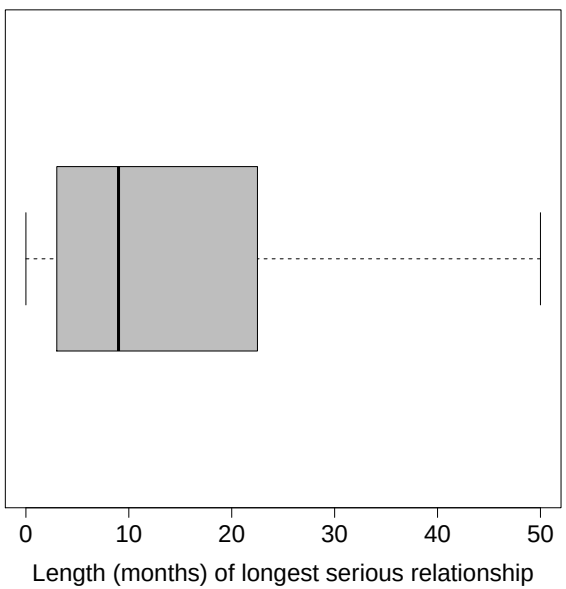


Which is true? (Don't use a calculator.)

1. Group 1 has the largest mean;
   Group 1 has largest standard deviation
2. Group 3 has the largest mean;
   Group 3 has largest standard deviation
3. all three groups have same mean;
   Group 1 has largest standard deviation
4. all three groups have same mean;
   Group 2 has largest standard deviation
5. all three groups have same mean;
   Group 3 has largest standard deviation

Group Exercise

```
 min Q1 median   Q3 max mean   sd  n missing
   0  3      9 22.5  50 12.5 12.3 63       4
```



Which of the following statements are true?

1. Exactly 50% of students had 9 months as their longest serious relationship
2. 50% of students had a longest serious relationship of 12.5 months or longer.
3. There are no students who have never been in a serious relationship.
4. 75% of students had serious relationships longer than 22.5 months.
5. None of these are true.

Summarizing and visualizing quantitative variables

Statistics:

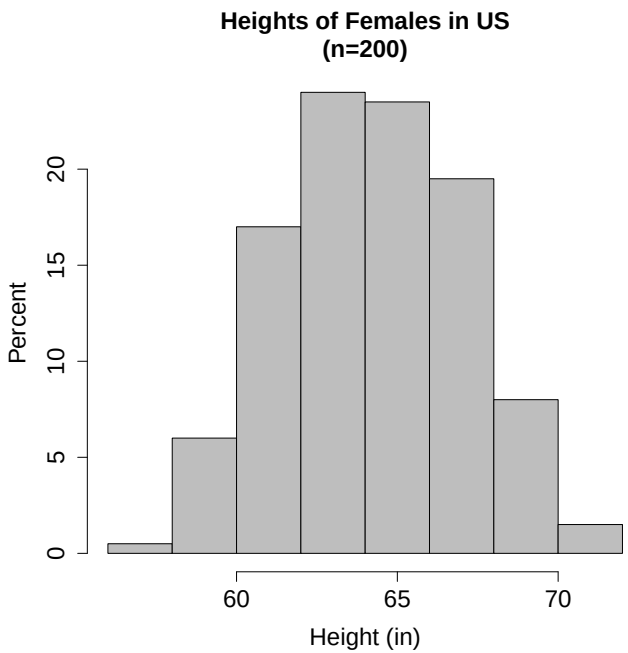
- Position: percentiles ($Q1 = 25^{th}$, median$= 50^{th}$, $Q3 = 75^{th} = Q3$)
- Center: mean, median
- Variability: standard deviation, interquartile range

◂ formulas for mean and sd ◂ finding percentiles and IQR

Figures:

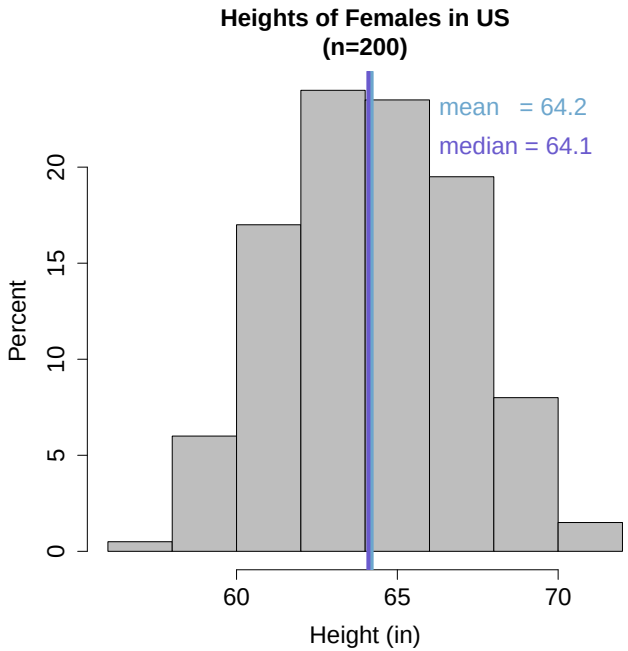
- dotplot - displays individual values
- histogram - displays values in bins
- boxplot - based on percentiles ◂ how to make a boxplot

Describing the Shape of Distribution



- **unimodal** - has one peak
- **symmetric** - mirror image when folded in half
- **bell-shaped** (normal) - data follow a bell-shaped curve

Mean vs Median (in symmetric data)

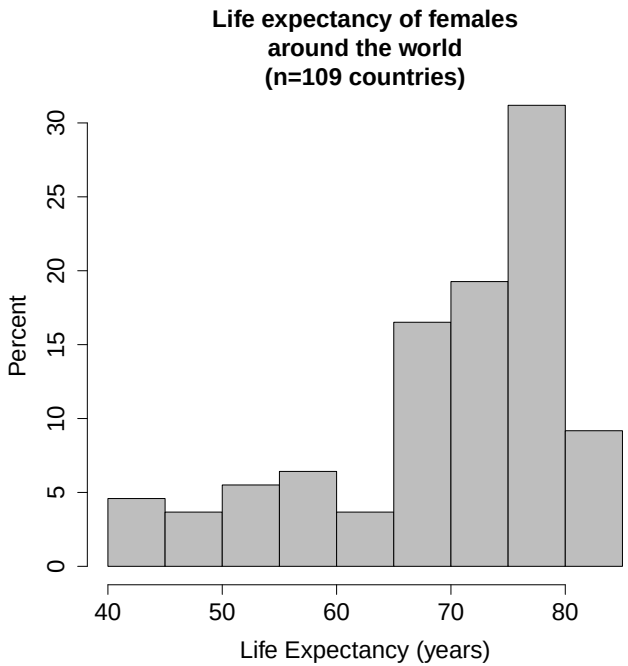


- For symmetric data, the mean and the median are approximately equal.
- In this case, the mean is an appropriate measure of central tendency.

If the mean and median are equal, this means that the data are bell-shaped.

1. True
2. False

Describing the Shape of Distribution

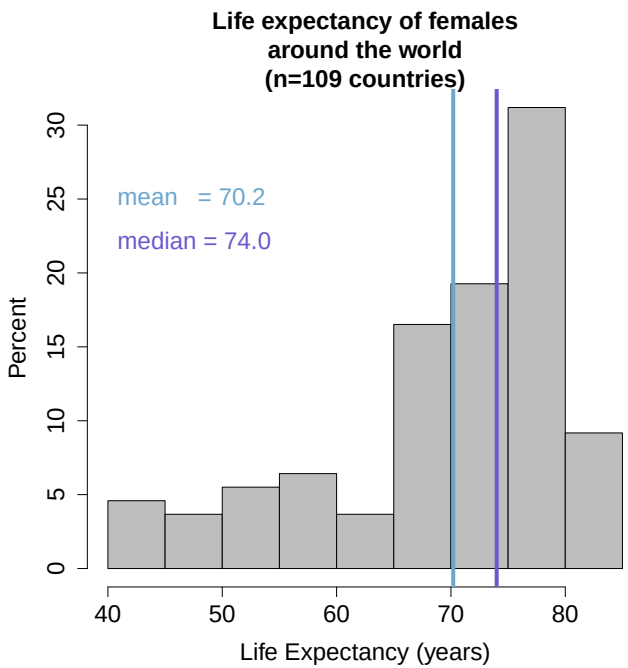


- **unimodal** - has one peak
- **left-skewed** - left tail is longer than the right (skew is in the direction of the tail)
- not symmetric
- not bell-shaped

Most countries have a life expectancy between 75 and 80 years.

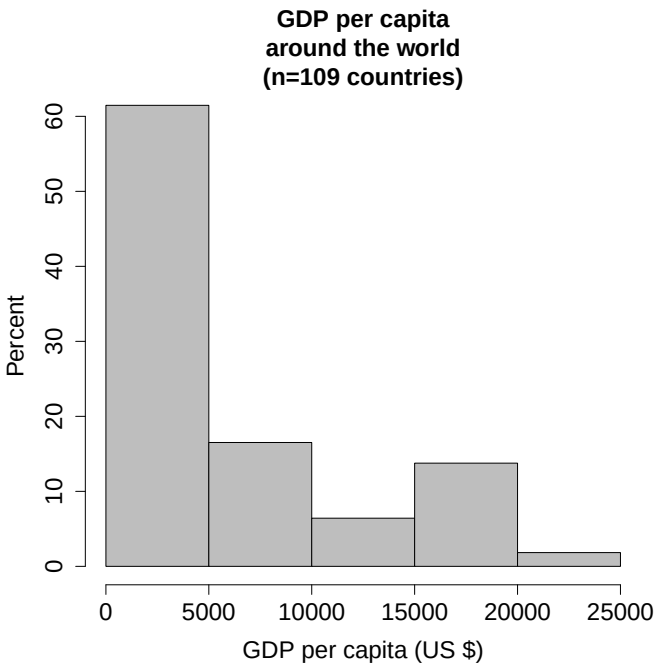
1. True
2. False

Mean vs Median (in left-skewed data)



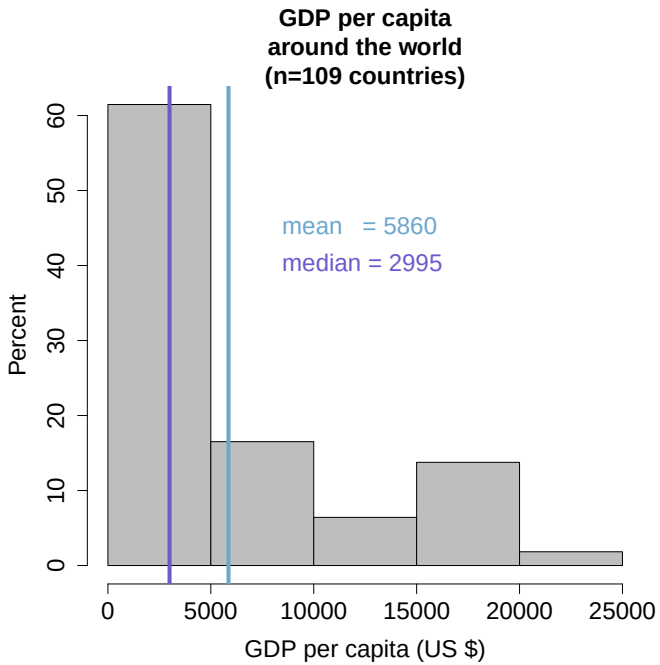
- For left-skewed data, the mean is less than the median.
- The mean is pulled in the direction of the long left tail.
- In highly skewed distributions, the median is preferred over the mean as a measure of central tendency (it better represents what is typical).

Describing the Shape of Distribution



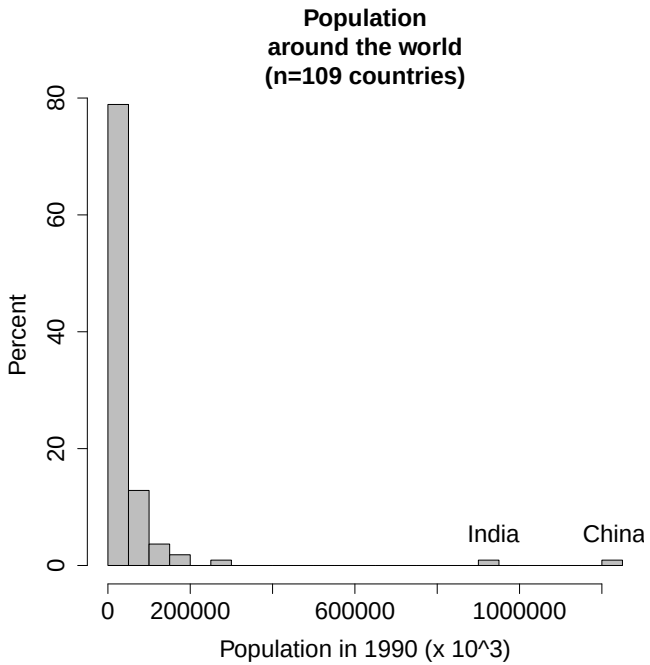
- **unimodal** - has one peak
- **right-skewed** - right tail is longer than the left (skew is in the direction of the tail)
- not symmetric
- not bell-shaped

Mean vs Median (in right-skewed data)



- For right-skewed data, the mean is greater than the median.
- The mean is pulled in the direction of the long right tail.
- In highly skewed distributions, the median is preferred over the mean as a measure of central tendency (it better represents what is typical).

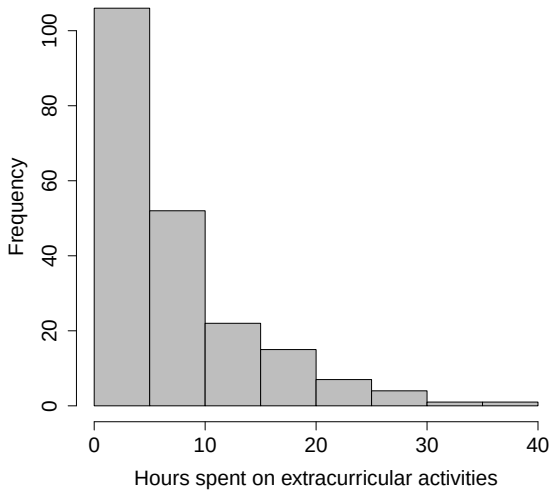
Describing the Shape of Distribution



- **unimodal** - has one peak
- **right-skewed** - right tail is longer than the left (skew is in the direction of the tail)
- has **outliers** - notice the gap between most of the observations and China and India
- not symmetric
- not bell-shaped

Group Exercise

208 students reported the typical weekly amount of time they spent on extracurricular activities (in hours).



Which of the following statements is *true*?

1. This distribution is left-skewed.
2. The mean is an appropriate measure of central tendency to represent a typical student response.
3. As the semester progresses, students are spending fewer hours on extracurricular activities.
4. The maximum hours spent weekly on extracurricular activities is greater than 100.
5. None of these statements are true.

Group Exercise

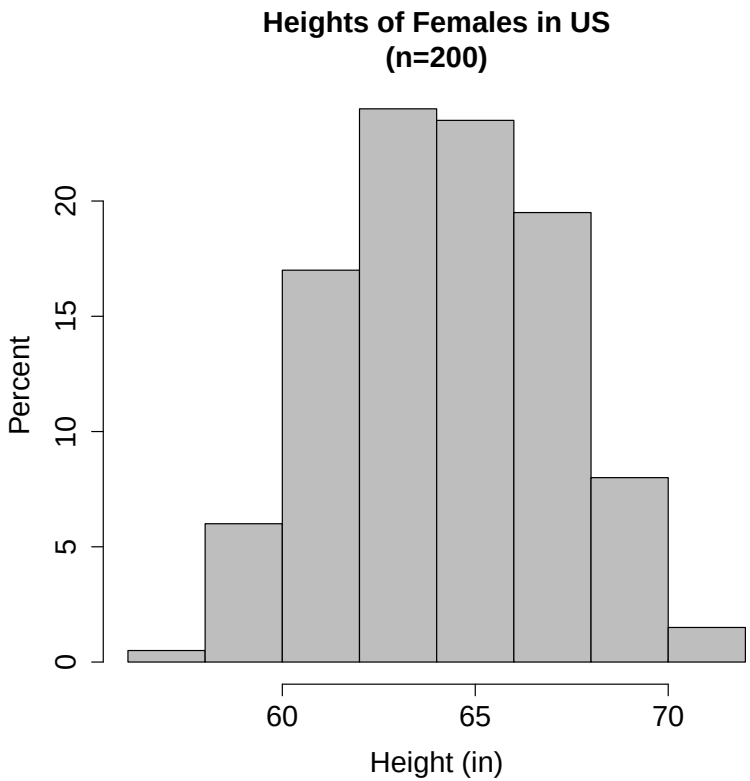
A real estate agent is trying to sell a house in a neighborhood in which most houses are worth \$180,000-\$220,000, but a few houses cost much more than that. The house for sale is listed at \$210,000, and the real estate agent is making the argument to the prospective home buyer that this is a really good deal because a typical house sells for \$250,000.

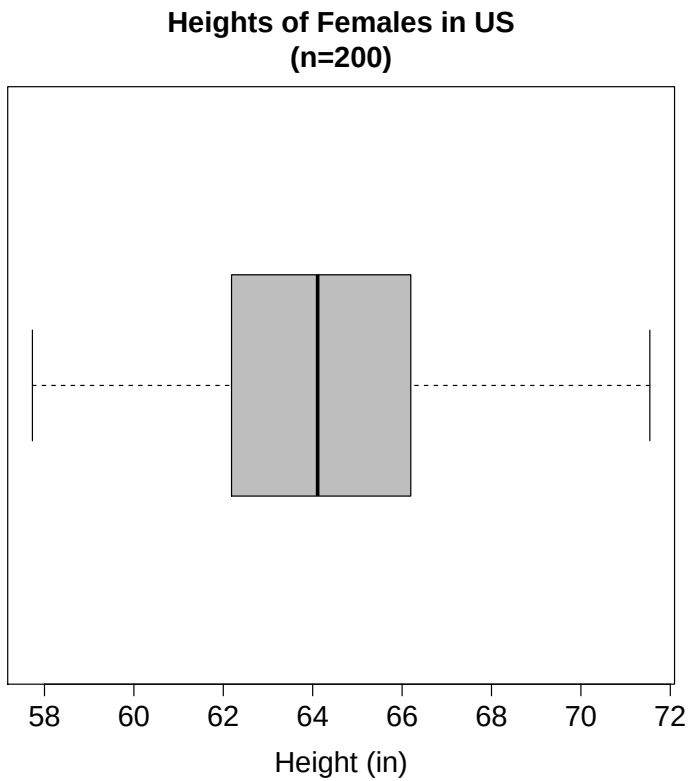
Which statistic is the real estate agent using to support her argument regarding the price of a 'typical' house?

1. the mean
2. the median
3. the mode
4. the standard deviation

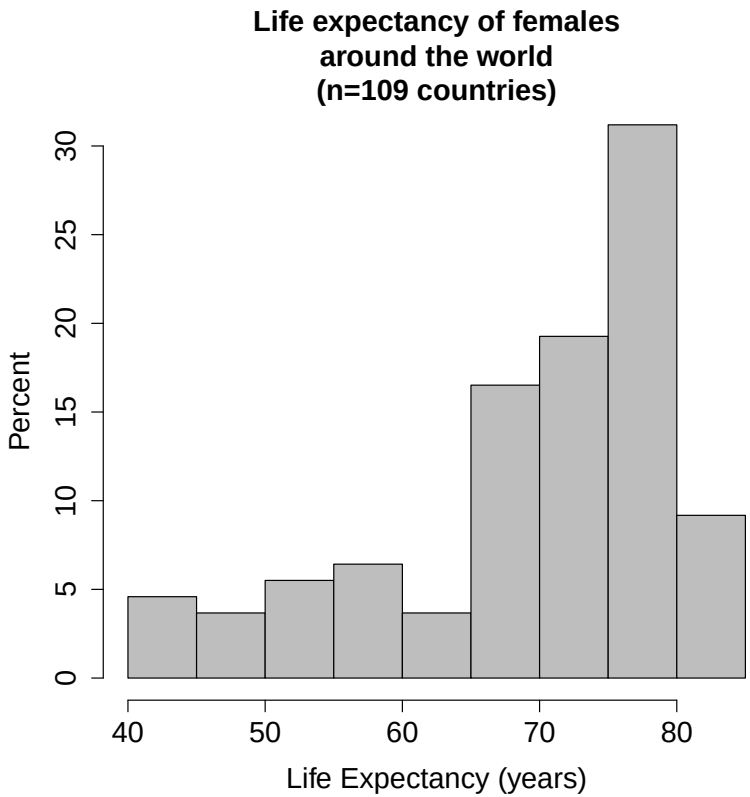
Is this a fair portrayal of 'typical' housing prices?

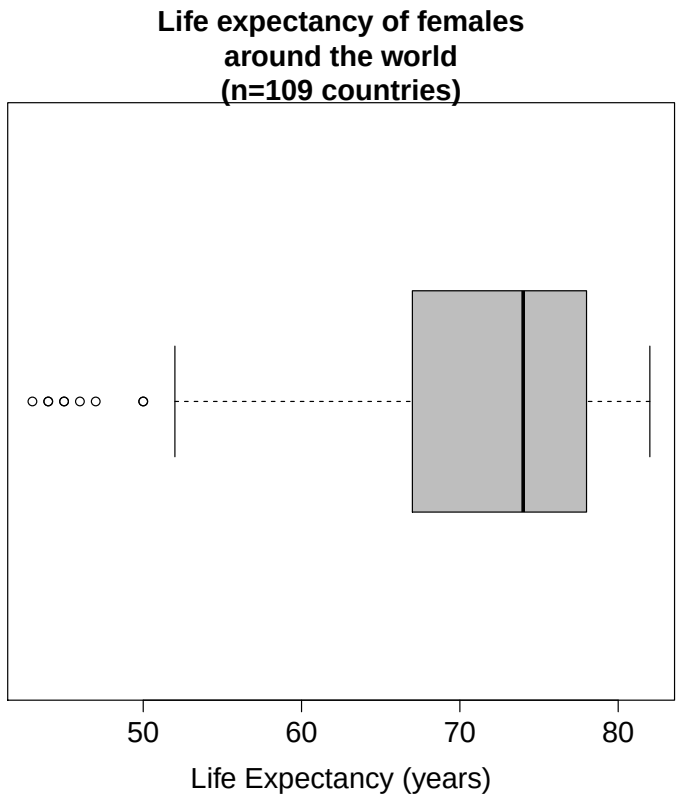
Histogram vs Boxplot



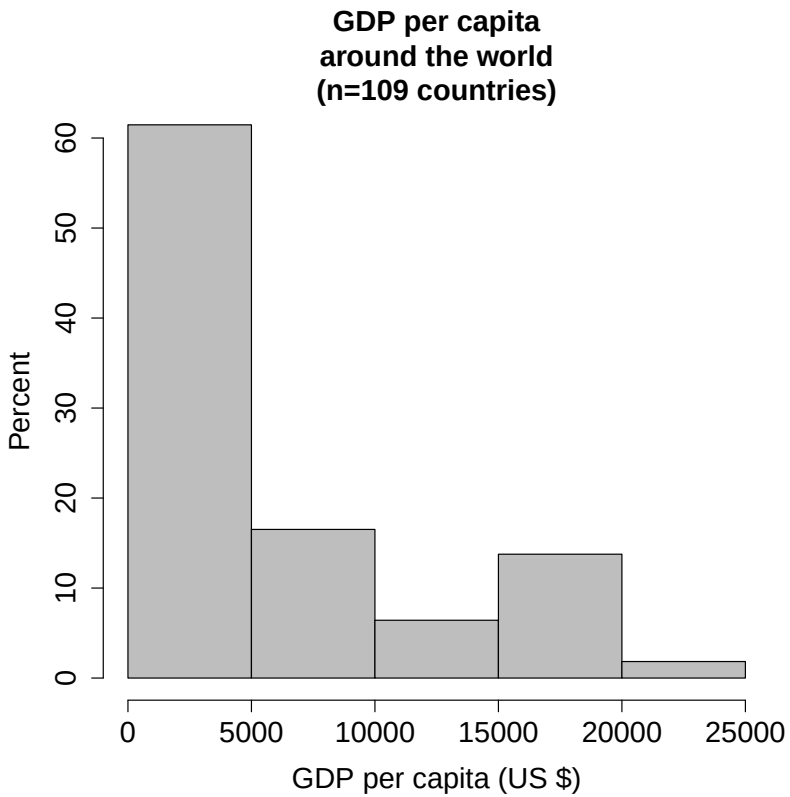


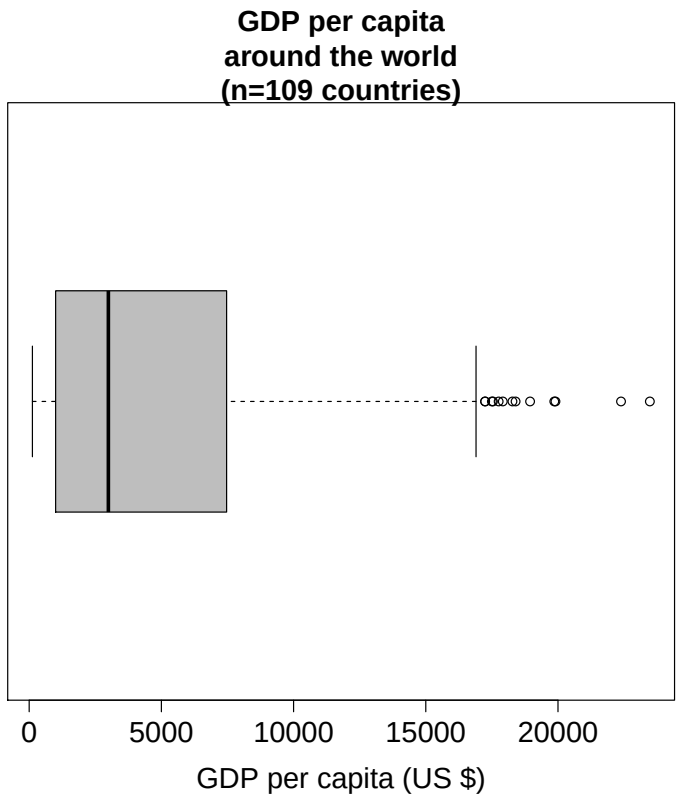
Histogram vs Boxplot



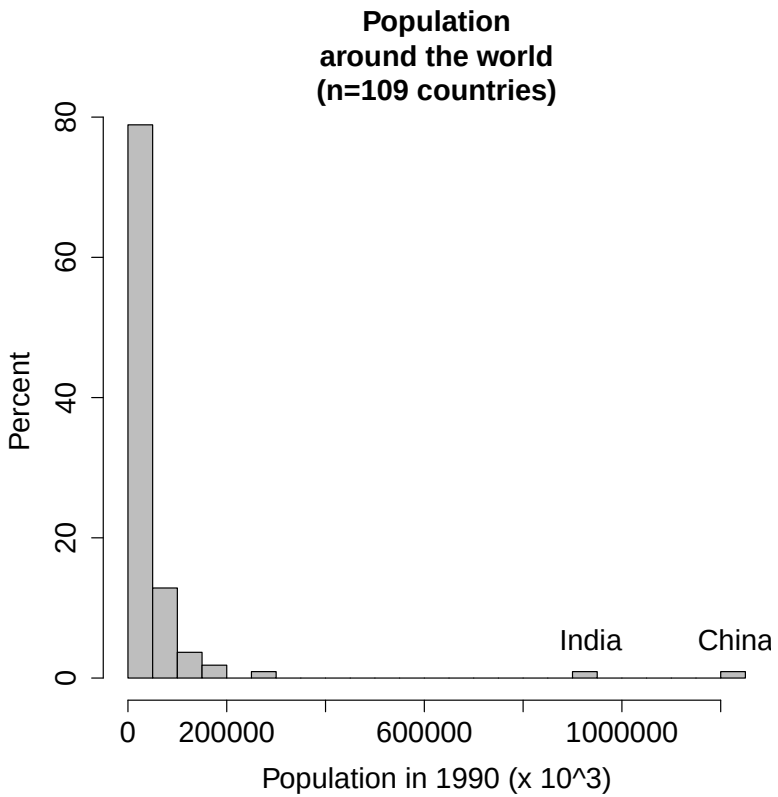


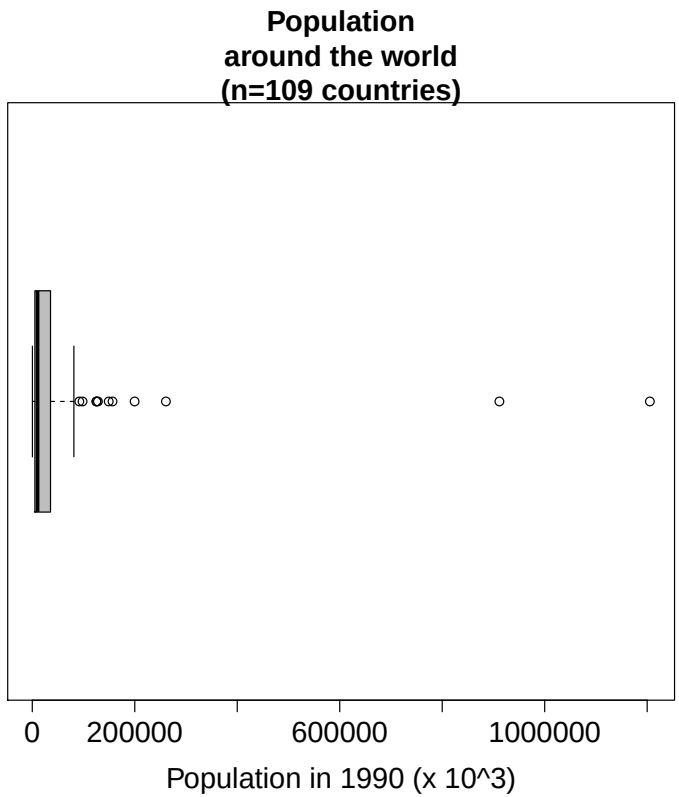
Histogram vs Boxplot





Histogram vs Boxplot





Group Exercise

This is a summary of the distribution of the number of hours spent weekly on extracurricular activities by 208 students.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	2.000	5.000	7.812	10.000	40.000

What is the most plausible shape of this distribution?

1. bell-shaped
2. right-skewed
3. left-skewed
4. none of these

Group Exercise

An experiment regarding the physiological cost of reproduction on male fruit flies contains the following variables. Male fruit flies were randomly assigned to cohabitate with one of 5 experimental groups of female fruit flies.

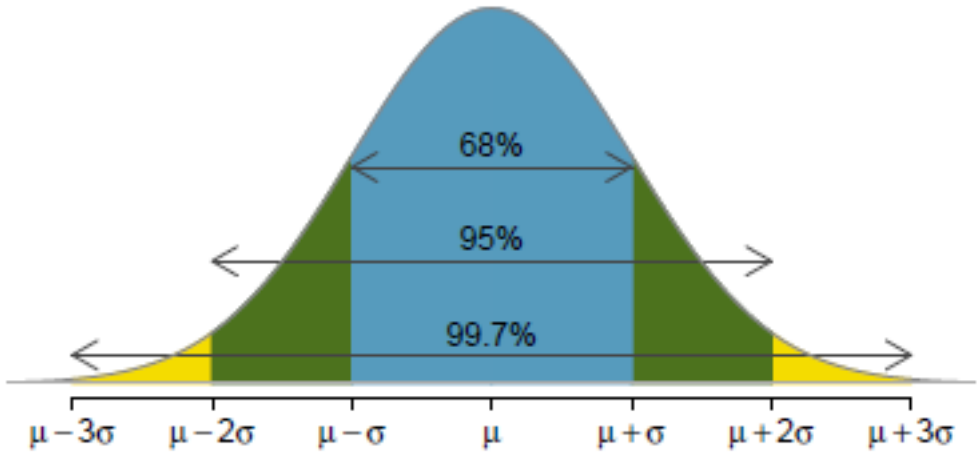
`type`	Type of experimental assignment 1 = no females 2 = 1 newly pregnant female 3 = 8 newly pregnant females 4 = 1 virgin female 5 = 8 virgin females
`lifespan`	lifespan (days)
`thorax`	length of thorax (mm)

Which figure would you use to plot `type`?

1. dotplot
2. histogram
3. bar plot
4. pie chart
5. boxplot

Normal distribution

When a distribution is *unimodal*, approximately *symmetric*, and *bell-shaped*, we describe it as a **normal** distribution..



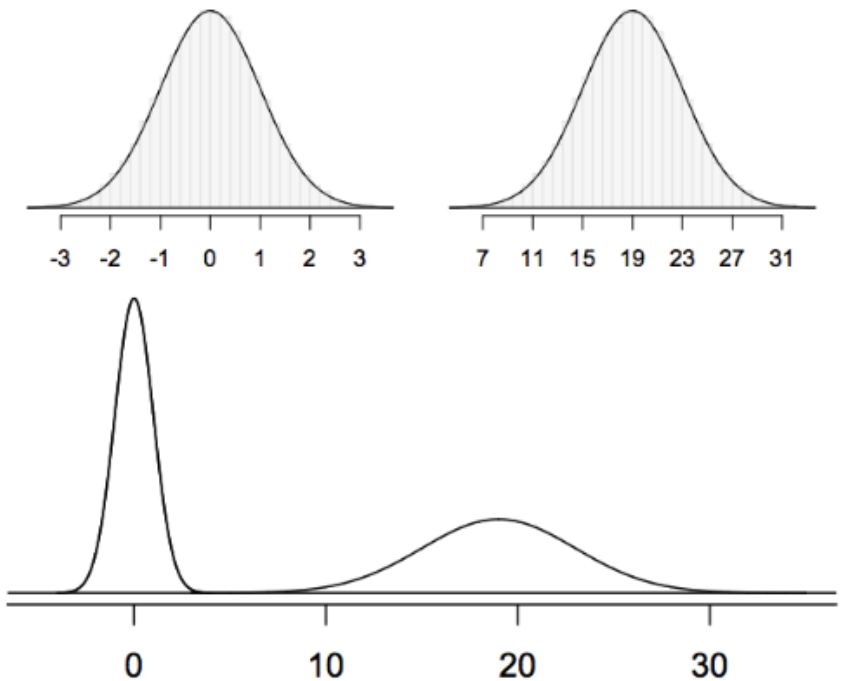
For any variable following a normal distribution

- 68% of observations fall within one standard deviation of the mean
- 95% of observations fall within two standard deviations of the mean
- 99.7% of observations fall within three standard deviations of the mean

Two normal distributions

μ: mean, σ: standard deviation

$N(\mu = 0, \sigma = 1)$ $N(\mu = 19, \sigma = 4)$



Using the normal distribution

68%	95%	99.7%
$\bar{x} \pm s$	$\bar{x} \pm 2s$	$\bar{x} \pm 3s$

Suppose women on average are 64 inches tall with a standard deviation of 3 inches. Sketch the distribution of heights of women.

- 68% of women are between ____and ____inches tall
- 95% of women are between ____and ____inches tall
- Nearly all (99.7%) women are between ____and ____inches tall
- About what percent of women are taller than 73 inches?

Group Exercise

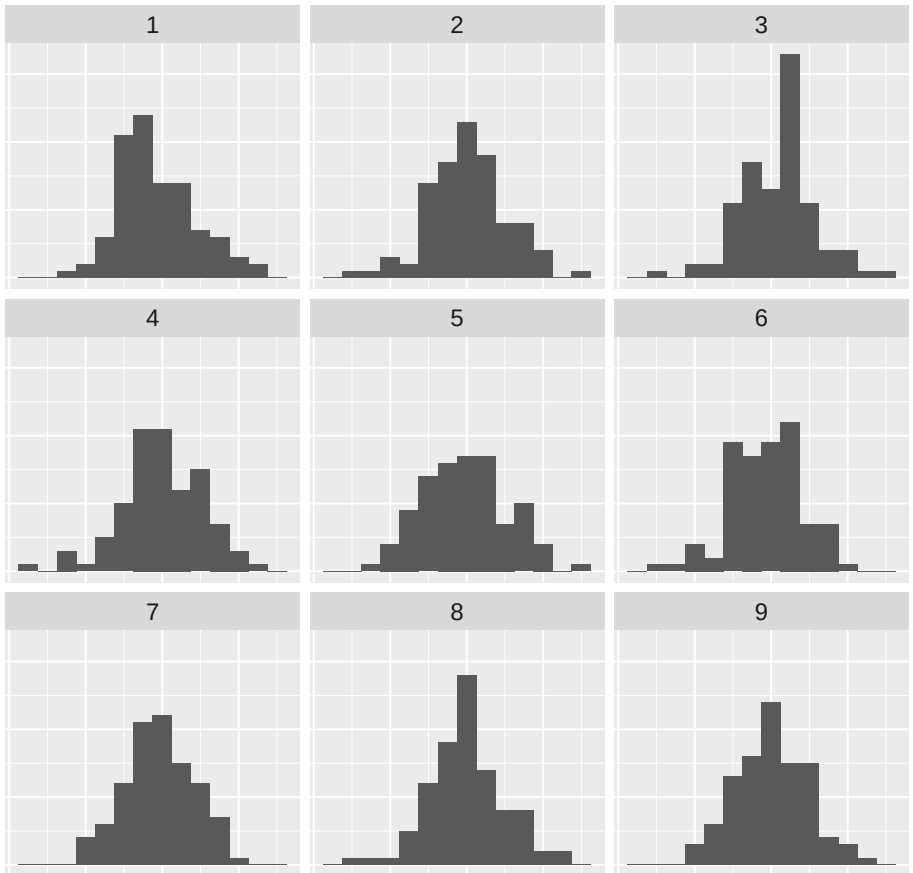
A doctor collects a large set of heart rate measurements that approximately follow a normal distribution. The doctor reports the the average heart rate is 110 beats per minute, the lowest is 65, and the highest is 155.

Which of the following is most likely to be the standard deviation of this distribution?

1. 5
2. 15
3. 35
4. 90

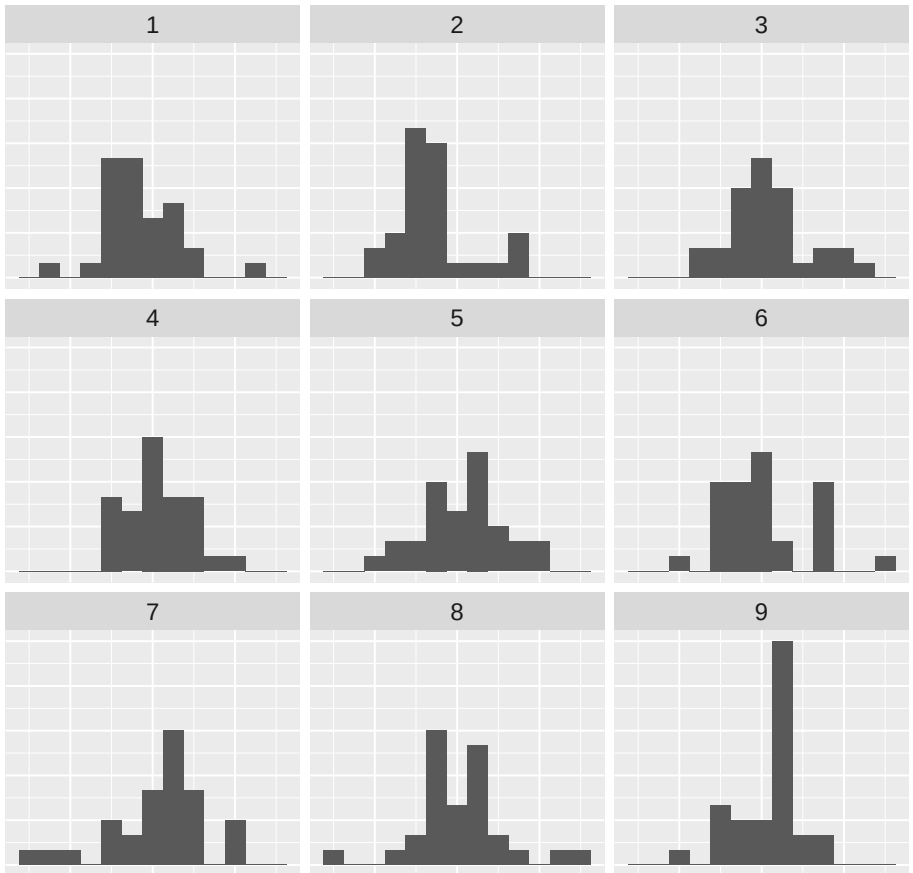
Group Exercise

Here we have 9 data sets from samples of size $n = 100$. Which of these 9 data sets come from a normal distribution?



Group Exercise

Here we have 9 data sets from samples of size $n = 30$. Which of these 9 data sets come from a normal distribution?

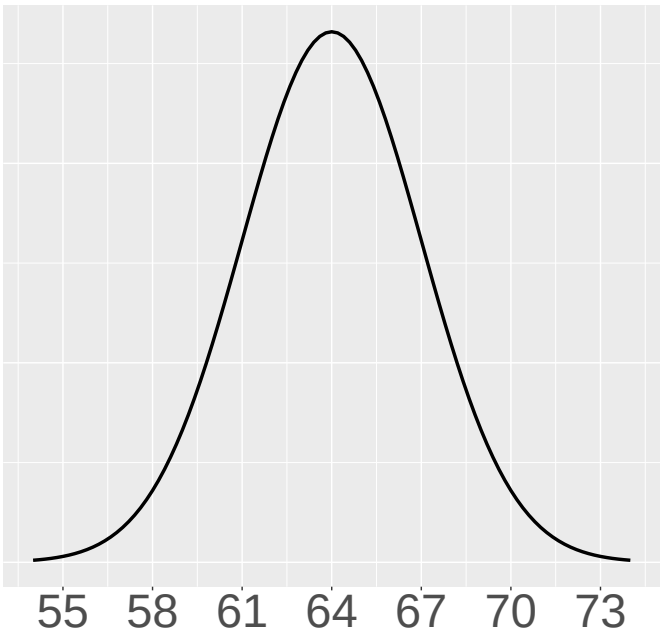


z-score

- Based on the normal distribution, we know it is unusual for an observation to fall more than three standard deviations away from the mean
- Therefore, one way we can assess if an observation is a potential outlier is to calculate **how many standard deviations away from the mean it is**.
- If an observation falls more than three standard deviations away from the mean, it can be regarded as a **potential outlier**.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

z-score example



mean= 64, sd= 3

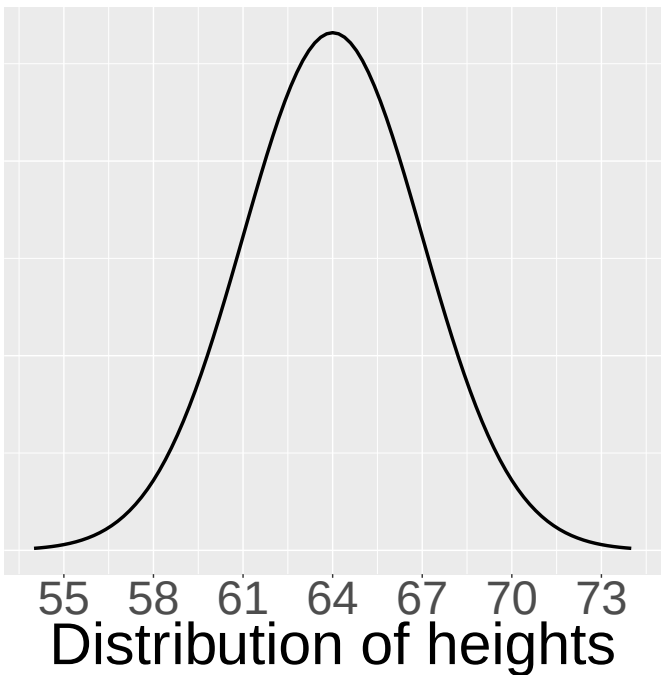
Suppose Mary is 67 inches tall.

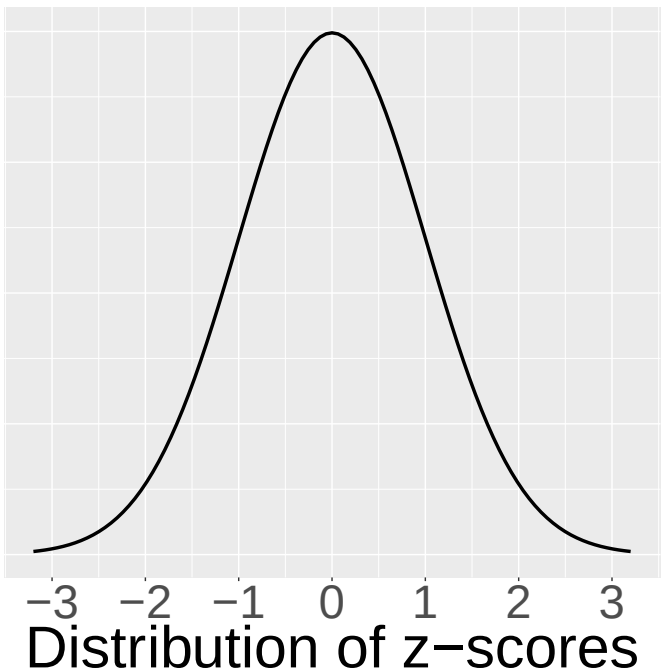
1. What is the z-score for Mary's height?

2. What is the interpretation of this z-score?

the distribution of z-scores

When a z-score is calculated from a normal distribution, the z-scores themselves follow a normal distribution with a mean of zero and a standard deviation of 1. We call this the standard normal distribution, and it is often referred as the z distribution.





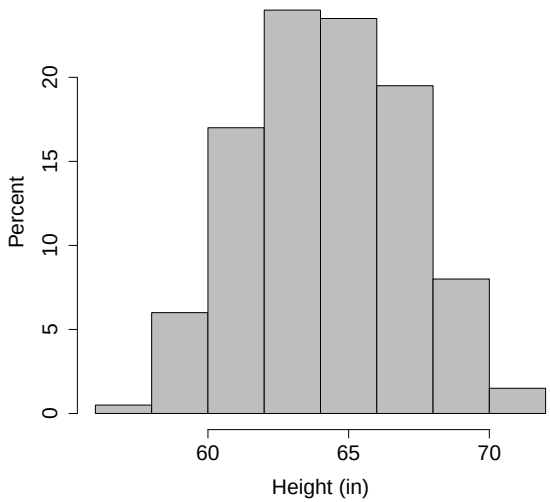
Group Exercise

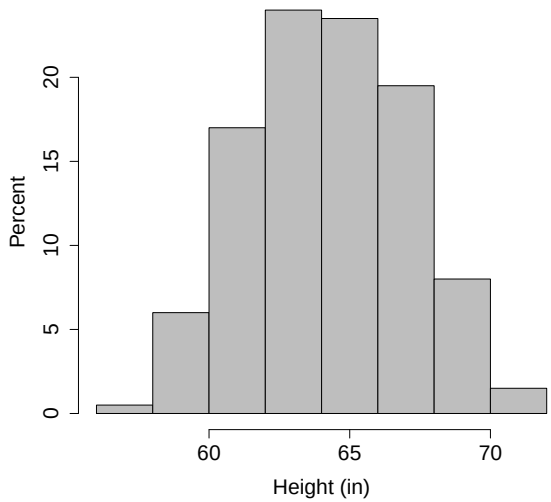
Suppose marketing and accounting majors have their own distribution of starting salaries (that is, each field has its own mean and standard deviation of salaries). Tom gets a job in marketing and Anna gets a job in accounting. Tom's z-score for his salary offer is 1.5, and Anna's is 0.67.

Which of the following is *true*?

1. Tom's salary offer was higher than Anna's.
2. Since Anna's z-score is less than 1, her salary offer was below the mean.
3. Anna's salary offer is relatively closer to the mean starting salary for her field than Tom's.
4. Tom's salary offer is 150% better than the mean starting salary for his field.
5. More than one statement is true.

Group Exercise





$n_1 = 100$

$\bar{x}_1 = 63.8$

$s_1 = ?$

$n_2 = 1,000$

$\bar{x}_2 = 63.8$

$s_2 = ?$

What is the relationship between s_1 and s_2?

1. $s_1 > s_2$
2. $s_1 < s_2$
3. $s_1 = s_2$

Mean and Standard Deviation

Mean (or average): the sum of the observations divided by the number of observations

$$\bar{x} = \frac{\sum x}{n}$$

The standard deviation represents a type of average distance of an observation from the mean.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Back

Percentiles

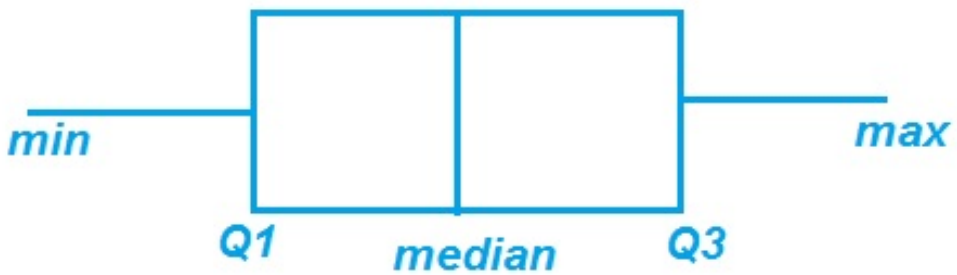
1. Order your data.
2. Identify the middle of the data. If n odd, the 50^{th} percentile (median) is the value in the middle. If n is even the 50^{th} percentile (median) is the average of the two middle values.
3. Examine the lower half of the data defined by the median (if n odd exclude median). The median of the lower half is the 25^{th} percentile (first quartile).
4. Examine the upper half of the data defined by the median (if n odd exclude median). The median of the upper half is the 75^{th} percentile (third quartile).

The interquartile range (IQR) of the data is the distance between the third and first quartiles: $IQR = Q3 - Q1$

Back

Boxplot

A **five-number summary** of data includes the minimum value, Q1, median, Q3, and maximum value. A five-number summary can be displayed in a **boxplot**.



The whiskers extend out to the smallest and largest observations that are **not** potential outliers. Potential outliers are indicated with circles. An observation is a **potential** outlier if

- it falls below $Q1 - 1.5 \times IQR$
- it falls above $Q3 + 1.5 \times IQR$